Creating lightweight ontologies for dataset description

Practical applications in a cross-domain research data management workflow

João Aguiar Castro
Faculdade de Engenharia da Universidade do Porto / INESC TEC
Portugal
joaoaguiarcastro@gmail.com

João Rocha da Silva
Faculdade de Engenharia da Universidade do Porto / INESC TEC
Portugal
joaorosilva@gmail.com

Cristina Ribeiro
DEI—Faculdade de Engenharia da Universidade do Porto / INESC TEC
Portugal
mcr@fe.up.pt

ABSTRACT

The description of data is a central task in research data management. Describing datasets requires deep knowledge of both the data and the data creation process to ensure adequate capture of their meaning and context. Metadata schemas are usually followed in resource description to enforce comprehensiveness and interoperability, but they can be hard to understand and adopt by researchers. We propose to address data description using ontologies, which can evolve easily, express semantics at different granularity levels and be directly used in system development. Considering that existing ontologies are often hard to use in a cross-domain research data management environment, we present an approach for creating lightweight ontologies to describe research data. We illustrate our process with two ontologies, and then use them as configuration parameters for Dendro, a software platform for research data management currently being developed at the University of Porto.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.3.5 [**Online Information Services**]: Data sharing

Keywords

Research data management, lightweight ontology, research data description

1. INTRODUCTION

Research data management comprises many complex challenges, and resource description stands out as a particularly hard one. Research domains are diverse in nature and comprise very specific concepts, making it necessary for researchers and data management professionals to work together in order to describe research datasets from a given domain. This is even more prevalent when targeting the so-called long-tail of research data—research datasets produced by countless small research groups without the financial resources to afford dedicated curation services for their data [1]. This problem is aggravated in the current context of massive data creation, particularly in research groups with access to limited resources [2].

It has been shown that the absence of timely description from the start of data production can yield lackluster descriptions [3]—a very practical example is when researchers leave their teams after publishing and have not described their datasets. A possible solution would be to have data curators accompany the research workflow—however, small research groups may struggle to keep up with the description demands posed by the existing datasets, and we may not expect researchers to spend much time in data description activities. A possible compromise scenario is to support researchers in the description of their data as they produce them, postponing curator intervention until later in the workflow.

To describe their datasets, researchers need to know what metadata to include in the descriptions, something usually prescribed by metadata schemas. However, these are often too complex in order to fulfil different metadata requirements [4], making the description process too time-consuming and diverting researchers from their main activities.

Data repositories are being created to preserve research data, but are in many cases targeted at "finished data", which is only available near the end of the research workflow. At our university, we are developing Dendro, a collaborative data management platform for small research groups. It is designed to support data description from the moment that data is created and uses Linked Open Data at the core. Its data model encourages data curators to model lightweight ontologies that can satisfy the description needs of each specific domain while retaining the interoperability characteristics of the ontology itself. This paper reports on ongoing research concerning lightweight ontology modelling for dataset description. After designing a generic lightweight ontology for research datasets, we instantiate the process in two engineering domains and claim that the constraints specified in our process can be generalised to any situation where datasets are organised as hierarchical structures of files and folders.

2. ONTOLOGIES FOR RESEARCH DATA

Working around the complexity of standardized metadata schemas, some authors have started to select sets of metadata descriptors suited for their particular application. This "mixing and matching" approach has yielded the notion of *Application Profile* [5]. However, even application profiles can be hard to understand and adopt; moreover, they are self-contained, meaning that they do not evolve incrementally and through reuse like ontologies. For datasets in a fast-paced, multi-domain research environment, a more incremental approach is desirable.

Ontologies have been presented as a possible solution for research data asset description. They satisfy all desirable metadata requirements [6] and are capable of representing the specific semantics of each research domain, while being able to evolve asynchronously. Yet, the flexibility allowed by ontology modelling processes and the attempts to model every aspect of each domain make it hard to use them in an actual cross-domain research data management environment. OBOE, an ontology for describing and synthesizing ecological observation data, is an example of a domain model whose concepts are very specialized [7]. Like many domain-specific ontologies, its modelling granularity is too fine for it to support a data management system. EXPO, an ontology for scientific experiments [8], is another case of correctness from a modelling perspective, but with a granularity making it unwieldy for usage in an operational data management workflow. Others like ESG [9] and CERIF [10] model cross-domain research concepts representing activities, entities or artifacts relevant for the research workflow that can be use for dataset description—for example *Experiment*, *Project* or *Participant*.

In the literature, ontologies with a large number of formal axioms and constraints are called "heavyweight ontologies", while those with a simpler approach are called "lightweight ontologies" [11, 12]. Dublin Core (DC)[1], for instance, is currently a widely used lightweight ontology, on par with FOAF (Friend of a Friend). Their simplicity and weak constraints make them easily processable by machines, and we have incorporated both directly in our proposed multi-domain research data management platform as sources of descriptors. By defining a limited number of classes, avoiding the definition of many object properties, and living out constraints and axioms, these ontologies become viable alternatives to support the data model of a research data management system, while being more easily manageable by curators.

Our work builds on past experience obtained from the implementation of a solution for collaborative research data management using Semantic MediaWiki [13]. We improve this solution by employing ontologies and a triple store, dispensing with a relational database—an approach also followed by a previous architecture designed for extensible, domain-agnostic scientific data management [14].

3. MODELING PROCESS

We want researchers to add description to data as they are created. The data curators have to provide sets of meaningfull descriptors, and to do this we propose the formalization of these descriptors through lightweight ontologies. Building up on past experience we recommend starting the process with a meeting between the curator (a data management expert) and the data creators (domain experts) [15]. During our meetings, we introduced basic data management notions—such as the concepts of metadata and descriptor. After that, we presented a set of descriptors from existing metadata schemas. These schemas were generic (such as Dublin Core) and research-oriented (such as EML or DDI). After validating a first set of descriptors, the researchers were asked to think about which information would be necessary to provide enough scientific context to allow others to verify, replicate, and reproduce the experiments from which the datasets were gathered [16]. After we selected domain-specific concepts to describe datasets, we searched for them in existing ontologies and in controlled vocabularies (for example the IEEE thesaurus). All the remaining vocabularies that could not be reused from existing vocabularies were introduced in a domain-specific namespace.

3.1 An extensible lightweight ontology

Many research datasets consist of sets of files and folders that researchers share within their groups. We have therefore designed a data management platform that uses ontologies to represent such directory structures and relate them to research-specific concepts such as *Experiment* or *Project*. The concepts covered in our *Research* ontology range from the level of the *research experiment* to the level of the *data file*. This means that we not attempt to represent the semantics of file contents (as is the case of VoID [17], for example), nor the organizational and administrative concepts at the *research project* level (these can be found, for instance, in CERIF). The three ontologies, however, complement each other: CERIF models highest-level organizational concepts (project-level), Research is targeted at the individual experiments, and VoID is a good approach for modelling the data produced in those experiments.

The `Research` ontology is therefore focused on representing metadata for *parts of a research project*. This means that, for instance, a *temperature measurement* stored in a file will not be represented with the ontology. But the ontology may include a temperature property to represent the temperature at which an experiment was conducted. An instance of this different (metadata) temperature property can be associated to the *File* or *Folder* of the experiment.

Part **1** of Figure 1 shows the complexity incurred in representing a *temperature* metadata value using two heavyweight ontologies (EXPO and OBOE). Such complexity is undesirable in an operational system despite its semantic richness, so we propose a simplified representation via a lightweight ontology (part **2**). We argue that, however, both approaches can co-exist: a dataset's metadata can be represented using lightweight ontologies in one system and then evolve, via ontology relations, to more heavyweight representations if the need should arise.

To model research concepts to match the *File-Folder* representation of datasets, we have created `Research`, a generic lightweight ontology. It serves as "extension point" from which other domain-specific ontologies can be derived in order to represent the descriptors required for each domain. It contains concepts such as *Experiment* or *Paper*, that can be reasonably mapped as *Files* or *Folders*. The assumption here is that the directory structure closely follows the different activities of a research project—for example, we consider the `Paper` concept to represent all the assets and activities

[1]We use here the OWL version available at `http://bloody-byte.net/rdf/dc_owl/`

in the process of creating a paper, and not the paper as *concrete entity* (unlike EXPO for example). When creating a lightweight ontology for experiments in a specific domain, the curator can also subclass `Experiment` to create specific types of experiments with their own properties, depending on the domain (see Figure 2).

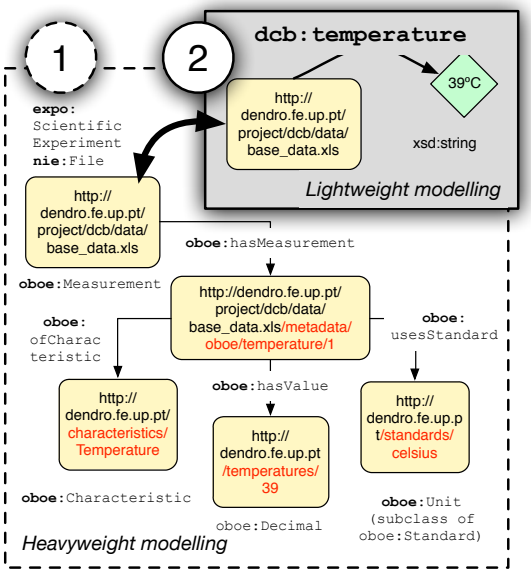

Figure 1: Recording dataset metadata using heavyweight and lightweight ontologies

3.2 Instantiating the process

In order to demonstrate the applicability of this process, we have considered two dataset case studies—one from fracture mechanics, and another concerning chemical analysis of pollutant concentrations.

- The Double Cantilever Beam (DCB), in fracture mechanics, consist in testing a given material to study its resistance. A specimen is subjected to pressure so that researchers can evaluate the propagation of cracks in it, recording force values applied and the corresponding specimen displacement. These values are then processed with appropriate methods and converted into energy values.

- Pollutant analysis (PA) is a type of experiment carried out by the analytical chemistry research group. This research group performs routine analysies regarding the concentration of certain pollutants in water and sediments collected *at a given time and place*, in what they call *runs*. These samples are taken and analysed using specific equipment and methods.

Figure 2 shows the lightweight ontologies proposed for the two case studies. The generic `Research` ontology is shown in **1**, the DCB-specific ontology is shown in **2** and the PA ontology is shown in **3**. The `DCBExperiment` is derived from `Experiment` to provide faceted representation (i.e. distinguishing the DCB datasets from the remainder). DCB experiments metadata must include the ambient `temperature` and `moisture` percentage at the location of the experiment, and the velocity at which the specimen was pressed (`testVelocity`). It is also important to record the specimen that was tested and its properties (`specimenLenght`, `specimenWidth`, `specimenHeight` and its `InitialCrackLenght`). They are subproperties of `specimenProperty` that can also be instantiated, allowing researchers to record other metadata.

Researchers from the pollutant analysis domain need to know the number of samples used (`sampleCount`) as well as the temporal and spatial information of the collected samples. To do so we used concepts from the Dublin Core ontology—namely `spatialCoverage`—to identify the place where the samples were taken, and specified `sampleCollectionDate` as a subproperty of DC `date`. Since this property is cross-domain, it was included in the `Research` ontology (**1**). Experiments are named *runs* by their creators, so `Run` was added as a subclass of `Experiment` and, as researchers compare `compound` values with legal limits, `LegislationDocument` (a `ResearchAsset` subclass) was created to represent relevant legislation (**3**).

4. CONCLUSIONS AND FUTURE WORK

In this paper we deal with the usage of ontologies in operational environments for the description of research datasets. We propose lightweight ontologies as a compromise between the specific requirements of data description for each domain and the working constraints of operational systems designed to manage datasets structured as file and folder hierarchies. We outline a lightweight ontology modelling process and demonstrate its applicability to a scenario where researchers describe their own datasets[2]. The descriptions are based on the properties defined in the ontologies.

Presently, we are actively developing Dendro, our ontology-based research data management platform. We will continue to apply this modelling process to create additional lightweight ontologies while working on improving the platform to make it increasingly suitable for researchers's tasks.

5. ACKNOWLEDGEMENTS

This work is supported by project NORTE-07-0124-FEDER-000059, financed by the North Portugal Regional Operational Programme (ON.2-O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT). João Rocha da Silva is also supported by research grant SFRH/BD/77092/2011, provided by the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

6. REFERENCES

[1] P. B. Heidorn, "Shedding Light on the Dark Data in the Long Tail of Science," *Library Trends*, vol. 57, no. 2, pp. 280–299, 2008.

[2] C. L. Borgman, J. C. Wallis, and N. Enyedy, "Little science confronts the data deluge: habitat ecology,

[2]Demonstrations of Dendro using these ontologies are available at `http://goo.gl/ug4FTh` and `http://goo.gl/SvdXhd`

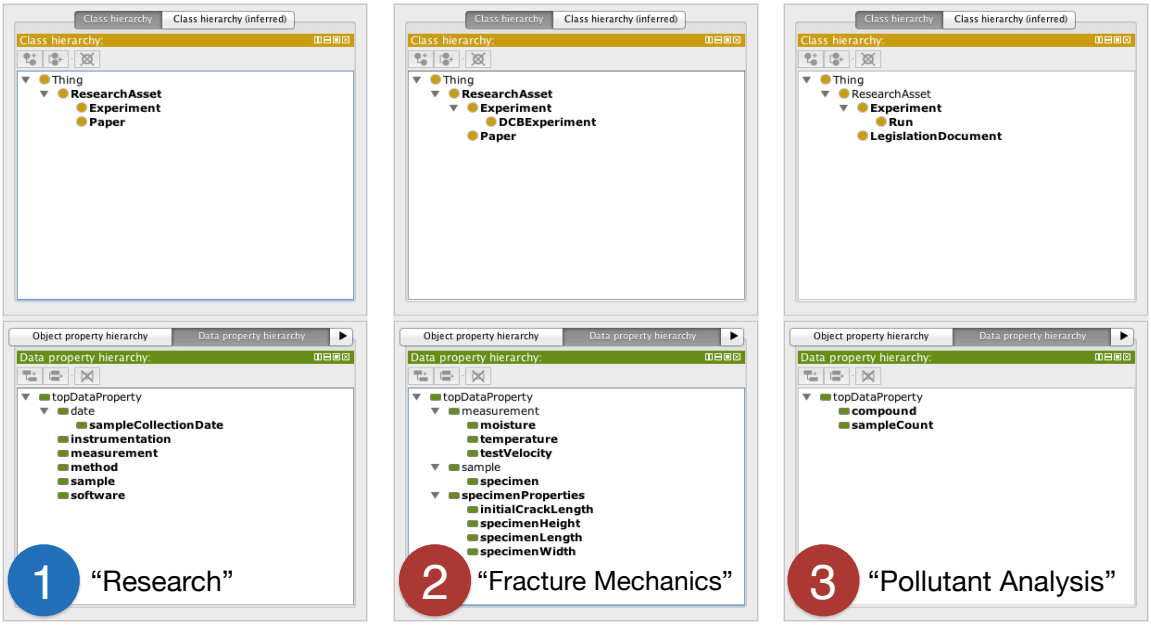

Figure 2: Our lightweight ontologies in Protégé—note the specialization of concepts from 1 in 2 and 3

embedded sensor networks, and digital libraries," *International Journal on Digital Libraries*, vol. 7, pp. 17–30, July 2007.

[3] L. Martinez-Uribe and S. Macdonald, "User engagement in research data curation," in *Proceedings of the 13th European conference on Research and advanced technology for digital libraries*, vol. 5714, pp. 309–314, Springer, 2009.

[4] J. Qin and K. Li, "How Portable Are the Metadata Standards for Scientific Data? A Proposal for a Metadata Infrastructure," in *Proceedings of the DC-2013 Conference*, pp. 25–34, 2013.

[5] R. Heery and M. Patel, "Application profiles: mixing and matching metadata schemas," *Ariadne*, no. 25, 2000.

[6] E. Duval, W. Hodgins, and S. Sutton, "Metadata principles and practicalities," *D-lib Magazine*, vol. 8, no. 4, pp. 1–10, 2002.

[7] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa, "An ontology for describing and synthesizing ecological observation data," *Ecological Informatics*, vol. 2, pp. 279–296, Oct. 2007.

[8] L. N. Soldatova and R. D. King, "An ontology of scientific experiments.," *Journal of the Royal Society, Interface / the Royal Society*, vol. 3, no. 11, pp. 795–803, 2006.

[9] L. Pouchard, L. Cinquini, and G. Strand, "The earth system Grid discovery and semantic web technologies," in *Workshop for Semantic Web Technologies for Searching and Retrieving Scientific Data - 2nd International Semantic Web Conference*, 2003.

[10] B. Jörg, "CERIF: The common European research information format model," *Data Science Journal*, vol. 9, no. July, pp. 24–31, 2010.

[11] O. Lassila and D. McGuinness, "The role of frame-based representation on the semantic web," *Linköping Electronic Articles in Computer and Information Science*, vol. 6, no. 005, 2001.

[12] O. Corcho, "Ontology based document annotation: trends and open research problems," *International Journal of Metadata, Semantics and Ontologies*, vol. 1, no. 1, pp. 47–57, 2006.

[13] J. Rocha, J. Barbosa, M. Gouveia, C. Ribeiro, and J. Correia Lopes, "UPBox and DataNotes: a collaborative data management environment for the long tail of research data," in *iPres 2013 Conference Proceedings*, 2013.

[14] Y.-f. Li, G. Kennedy, F. Ngoran, and P. Wu, "An Ontology-centric Architecture for Extensible Scientific Data Management Systems," *Future Generation Computer Systems*, vol. 29, no. 2, pp. 1–38, 2013.

[15] J. A. Castro, J. a. R. da Silva, and C. Ribeiro, "Designing an Application Profile Using Qualified Dublin Core: A Case Study with Fracture Mechanics Datasets," *Proc. of the International Conference on Dublin Core and Metadata Applications*, pp. 47–52, 2013.

[16] J. Greenberg and C. Hill, "Functional and Architectural Requirements for Metadata : Supporting Discovery and Management of Scientific Data," in *Proceedings of the DC-2012 Conference*, pp. 62–71, 2012.

[17] K. Alexander and M. Hausenblas, "Describing linked datasets-on the design and usage of void, the vocabulary of interlinked datasets," *In Linked Data on the Web Workshop (LDOW 09)*, 2009.